Comparison for alternative imputation methods for ordinal data

Federica Cugnata e Silvia Salini

DEMM, Università degli Studi di Milano

22 maggio 2013

Missing imputation

- Approaches:
  - A. Create a **complete dataset** (complete-case analysis or listwise deletion, available-case analysis, weighting procedures, imputation-based procedures)
  - B. **Model-based procedures** (inferences are based on likelihood or Bayesian analysis)
- Procedures:
  - A. **Univariate**, that substantially use information from the distribution of the variable with missing itself (i.e mean, median, mode, random imputation, ect)
  - B. **Multivariate**, that use the observed pattern for one or more related variables to estimate trough a model the variable with missing (i.e. linear and non linear regression models).
- Methods:
  - A. **Single imputation** (SI) that impute one value for each missing item;
  - B. **Multiple imputation** (MI) that impute more than one value for each missing item, to allow appropriate assessment of imputation uncertainty.

- Ratings are interpreted as the result of a cognitive process, where the judgement is intrinsically continuous but it is expressed in a discrete way within a prefixed scale of m categories.
- Final choices of respondents is the result of two components: a personal *feeling* and some intrinsic *uncertainty* in choosing the ordinal value of the response.

- The first component is expressed by a shifted Binomial random variable.
- The second component is expressed by a Uniform random variable.
- The two components are linearly combined in a mixture distribution.

$$P_r(R=r) = \pi \binom{m-1}{r-1} \xi^{m-r}(1-\xi)^{r-1} + (1-\pi)\frac{1}{m}, \qquad r = 1, 2, \ldots, m.$$

The parameters $\pi \in (0, 1]$ and $\xi \in [0, 1]$, and the model is well defined for a given $m > 3$. The acronym CUB stands for a Combination of Uniform and (shifted) Binomial random variables.

General formulation of a CUB(p,q) model with p covariates to explain uncertainty and q covariates to explain feeling:

1. A *stochastic component*:

$$Pr(R_i = r \mid \mathbf{y}_i;\, \mathbf{w}_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right),$$

   for r=1,2,...,m and for any $i = 1, 2, \ldots, n$.

2. Two *systematic components*:

$$\pi_i = \frac{1}{1+e^{-\mathbf{y}_i \boldsymbol{\beta}}}; \qquad \xi_i = \frac{1}{1+e^{-\mathbf{w}_i \boldsymbol{\gamma}}}; \qquad i = 1, 2, \ldots, n,$$

   where $\mathbf{y}_i$ and $\mathbf{w}_i$ denote the covariates of the i-th subject, selected to explain π_i and ξ_i, respectively.

$\mathbf{X}^{obs}$ and $\mathbf{X}^{mis}$ matrices with the covariates corresponding to the observed and missing cells of $\mathbf{x}$, respectively.
Based on the subset of the complete data, estimate the CUB model

$$Pr(\mathbf{x}^{i \in obs} = r \mid \mathbf{X}^{i \in obs}) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right)$$

$r = 1, 2, \ldots, m.$

$$\pi_i = \frac{1}{1+e^{-\mathbf{X}^i \beta}}\,; \qquad \xi_i = \frac{1}{1+e^{-\mathbf{X}^i \gamma}}\,; \qquad i \in obs\,,$$

To obtain a more efficient method it is possible to use a stepwise strategy to select only the significant covariates and estimate the best CUB model.

Each missing value $\mathbf{x}^{mis}$ is replaced by the mode of s random numbers from the estimated CUB model. To simulate from the given CUB distribution we use the inverse transform method. We generate a random number U and transform U into $\mathbf{x}$ as follows

$$\mathbf{x} = r \quad if \quad \sum_{j=1}^{r-1} p_j \leq U < \sum_{j=1}^{r} p_j$$

When more than one variable has missing data, imputation typically requires an iterative method of repeated imputations. On the basis of the Iterative Robust Model-based Imputation (**Templ et. al 2011)**

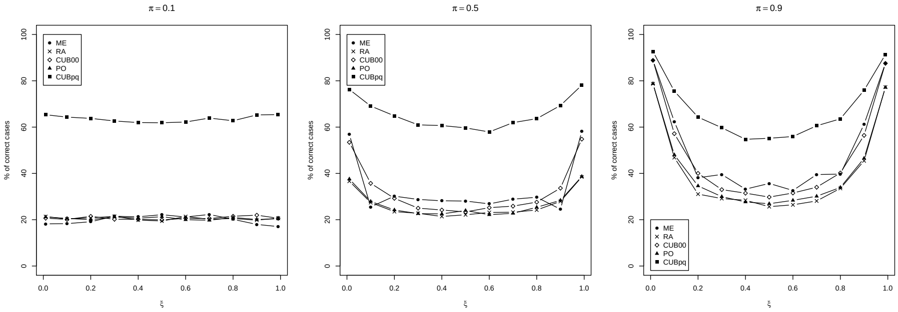
	Overall	Recommend	Repurchasing	Training	Software
CCA	0.501	0.537	0.552	0.657	0.388
ACA	0.582	0.625	0.637	0.695	0.391
ME	0.586	0.631	0.64	0.803	0.291
	(0.035)	(0.034)	(0.034)	(0.028)	(0.032)
RA	0.581	0.626	0.635	0.729	0.404
	(0.035)	(0.034)	(0.034)	(0.031)	(0.034)
CUB00	0.581	0.621	0.64	0.655	0.399
	(0.035)	(0.034)	(0.034)	(0.033)	(0.034)
FO	0.581	0.631	0.64	0.714	0.384
	(0.035)	(0.034)	(0.034)	(0.032)	(0.034)
MF	0.581	0.631	0.64	0.739	0.315
	(0.035)	(0.034)	(0.034)	(0.031)	(0.033)
CUBpq	0.586	0.631	0.64	0.69	0.401
	(0.035)	(0.034)	(0.034)	(0.032)	(0.035)

Benchmarking Analysis, Mattei et. al 2012

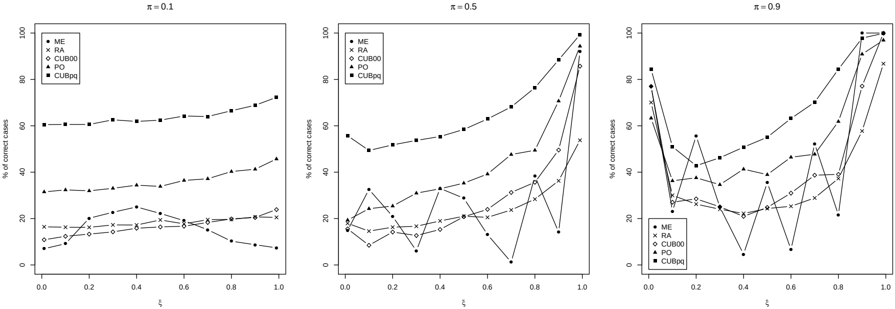
	Overall	Recommend	Repurchasing	Training	Software
CCA	0.501	0.537	0.552	0.657	0.388
ACA	0.582	0.625	0.637	0.695	0.391
MRA	0.582	0.629	0.636	0.715	0.393
	(0.035)	(0.034)	(0.034)	(0.04)	(0.038)
MCUB00	0.582	0.623	0.639	0.681	0.393
	(0.035)	(0.034)	(0.034)	(0.042)	(0.04)
MPO	0.583	0.625	0.635	0.696	0.368
	(0.035)	(0.034)	(0.034)	(0.034)	(0.036)
MLI	0.581	0.625	0.636	0.664	0.373
	(0.035)	(0.034)	(0.034)	(0.037)	(0.036)
MCUBpq	0.581	0.627	0.637	0.678	0.386
	(0.035)	(0.034)	(0.034)	(0.042)	(0.037)

- Imputation for one variable with covariates
- Variable Y is generated from CUB(0,0) with $m = 5$ for different values of the CUB parameters: $\pi = 0.1, 0.2, 0.3, \dots, 1$ and $\xi = 0, 0.1, 0.2, 0.3, \dots, 0.99$.
- Two covariates are generated, $X_1 \sim N(y, 0.16)$ and $X_2 \sim \texttt{CUB}(0, 1)$ with Y as covariate of the feeling.
- The 100 missing values are assigned a) randomly, b) when Y assumes low values and c) when Y assumes high values.
- **Mattei et al. (2012)**: polytomous regression (PO)

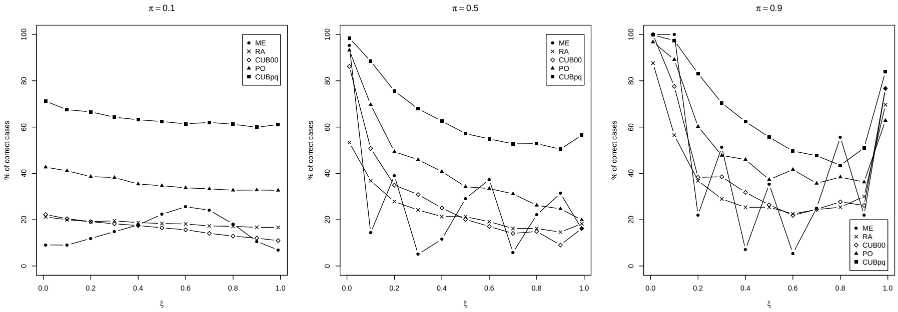
Imputation for one variable with covariates. % of correct cases, case (A)



Imputation for one variable with covariates. % of correct cases, case (B)



Imputation for one variable with covariates. % of correct cases, case (C)



- Missing values are present to more than one ordinal variables
- We considers two cases:
  - A) Variables $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ is a multivariate normal distribution where $Y_i \sim N(0,1)$ and $\rho(Y_i, Y_j) = 0.3, 0.5, 0.8$. Numerical values of Y are transformed into ordinal categories (Likert scale) using $m = 5$.
  - B) A variable W is generated from `CUB(0,0)` with $m = 5$;
    $\pi = 0.1, 0.2, 0.3, \ldots, 1$ and $\xi = 0, 0.1, 0.2, 0.3, \ldots, 0.99$.
    Y_1, Y_2, Y_3, Y_4, Y_5 are generated from `CUB(0,1)` with W as covariate of the feeling.
- **Ferrari et al. (2011)**: forward imputation (FO); **Stekhoven and Bühlmann (2012)** missForest (MF)

Results, Case A - % correct cases

Method	ρ 0.3	0.5	0.8
ME	24.17	23.2	23.855
	(2.534)	(2.861)	(2.801)
RA	20.109	20.158	19.067
	(2.717)	(2.093)	(2.735)
CUB00	20.703	20.933	20.933
	(2.539)	(2.3)	(2.127)
PO	23.539	28.133	41.855
	(3.013)	(2.819)	(3.746)
FO	23.515	28.461	42.097
	(2.271)	(2.153)	(2.868)
MF	25.37	30.17	**48.933**
	(3.072)	(2.956)	(3.028)
CUBpq	**26.158**	**32.473**	46.352
	(2.782)	(3.376)	(3.274)

Simulation Study 2

Results, Case B - % correct cases

	$\pi = 0.1$			$\pi = 0.5$			$\pi = 0.9$		
				ξ					
Method	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
ME	24.48	25.76	24.4	46.48	29.76	28.16	54.32	31.12	30.4
	(1.11)	(1.94)	(2.37)	(4.07)	(2.69)	(0.742)	(0.64)	(3.14)	(3.24)
RA	28.32	29.92	26.24	36.24	29.2	26.88	41.76	28.88	27.92
	(1.98)	(3.61)	(3.02)	(2.03)	(1.49)	(2.51)	(2.70)	(2.29)	(1.52)
CUB00	35.52	33.44	34.88	44.48	36.56	31.68	53.44	32.08	34.16
	(4.13)	(3.34)	(1.70)	(1.41)	(1.70)	(3.02)	(1.17)	(1.647)	(2.08)
PO	39.92	42	42.24	46.4	41.2	40.48	47.84	36.72	37.68
	(4.73)	(2.49)	(2.21)	(1.53)	(3.51)	(3.09)	(3.45)	(1.87)	(2.42)
FO	42.64	44.24	43.36	44.4	40.64	41.04	49.28	38.64	38.96
	(2.66)	(3.24)	(2.17)	(1.71)	(1.83)	(2.49)	(4.59)	(0.69)	(2.21)
MF	41.76	46.4	45.6	46.32	42.56	40.56	50.88	39.76	41.92
	(2.35)	(2.69)	(3.39)	(1.74)	(3.18)	(2.73)	(2.39)	(0.89)	(2.69)
CUBpq	**49.04**	**51.04**	**52.16**	**55.04**	**47.68**	**48.32**	**54.88**	**44.4**	**44.72**
	(2.03)	(2.00)	(2.85)	(3.66)	(2.47)	(4.72)	(1.56)	(1.97)	(2.57)

D'Elia and Piccolo, 2005: *Emergency in Metropolitan Area*

Variable	π	ξ
Political Patronage	0.63	0.70
Organized Crime	0.90	0.94
Unemployment	0.67	0.77
Pollution	0.75	0.24
Public Health	0.72	0.51
Petty Crimes	0.77	0.74
Immigration	0.57	0.04
Street and Waste	0.66	0.29
Traffic Transport	0.82	0.20

- Wave 2006, 419 observation
- missing cases 10%:
  - A) missing at random
  - B) missing on the low categories
  - C) missing on the high categories
  - D) missing associated to some values of the covariates

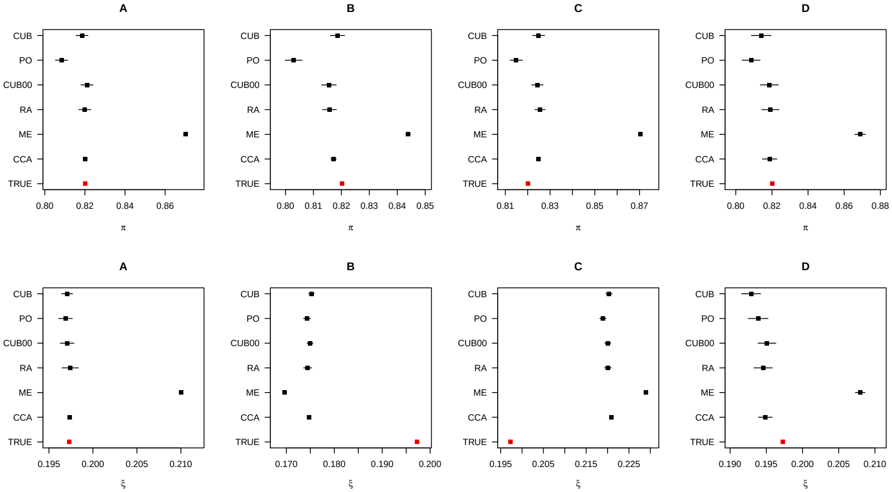
Application 1

Results: % of correct cases

		ME	RA	CUB00	PO	CUBpq
Case A	1	19.55	16.55	18.57	16.41	**20.18**
	7	15.11	28.66	**41.73**	31.91	39.87
	9	20.31	18.61	22.36	19.62	**22.77**
Case B	1	4.82	17.48	20.22	17.34	**20.64**
	7	0.00	8.91	2.94	**11.26**	9.38
	9	0.00	12.59	10.18	**14.17**	11.51
Case C	1	0.00	11.29	**13.12**	10.03	12.95
	7	2.74	42.23	**66.33**	51.88	63.23
	9	3.07	21.40	**26.24**	22.61	24.71
Case D	1	14.50	15.25	17.89	17.13	**18.41**
	7	12.77	32.78	49.60	38.22	**52.70**
	9	17.58	17.38	20.45	**21.98**	21.37

Application 1

Results: estimation of the CUB parameters, *Political Patronage*



Application 2

Airline Industry

Typical questionnaire filled by passengers of airlines companies to evaluate flight. Seven points scale (from 1 = extremely dissatisfied to 7 = extremely satisfied). Covariates related to the flight and covariates related to the passenger are present

Variable	π	ξ
overall booking	0.68	0.19
check-in	0.46	0.15
departure	0.82	0.24
cabin environment	0.94	0.25
meal	0.97	0.27

$n = 558$, missing cases 10%:

A) missing at random

B) missing on the low categories

C) missing on the high categories

D) missing associated to some values of the covariates

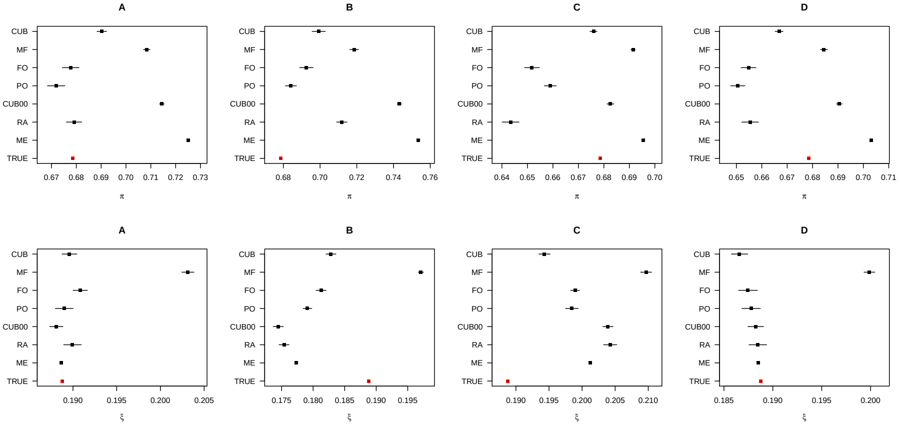
Application 2

Results: % of correct cases

		ME	RA	CUB00	PO	FO	MF	CUBpq
Case A	1	41.85	24.55	32.17	31.37	29.48	**35.14**	33.06
	2	34.88	20.84	24.96	36.94	34.53	32.56	**38.31**
	3	39.61	23.40	32.88	34.13	42.38	**43.90**	35.74
Case B	1	21.61	14.82	17.99	16.73	18.01	**22.98**	19.79
	2	10.04	10.94	18.86	20.80	18.27	20.98	**21.90**
	3	8.04	15.09	16.38	24.94	22.59	**29.59**	21.88
Case C	1	62.94	32.25	43.19	40.42	41.48	43.66	**49.61**
	2	0.00	26.79	34.70	47.45	42.27	42.97	**52.39**
	3	0.00	30.76	38.40	42.56	50.95	**52.01**	48.78
Case D	1	48.03	28.81	34.33	32.11	34.09	**39.81**	39.53
	2	31.09	21.67	25.96	36.89	36.78	37.34	**39.59**
	3	31.35	23.18	31.96	35.80	40.75	**46.57**	37.42

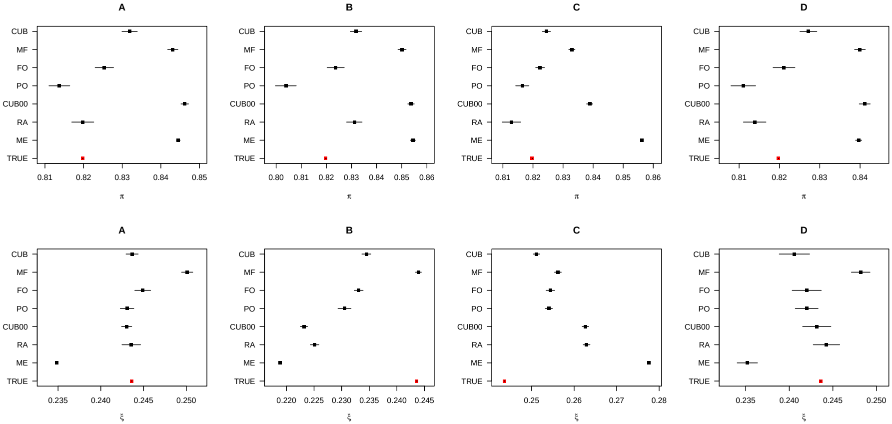
Application 2

Results: estimation of the CUB parameters *Overall booking*



Application 2

Results: estimation of the CUB parameters *Check-in*



Conclusions

Imputation for missing ordinal data is more complex than for continuous data. Literature proposals: *forward imputation* (FO) and *missForest* (MF). When `CUB` is the model of analysis of the data that you want to use, a further opportunity is to use `CUB` also for imputation (complete-case and model-based optical). Simulation studies and test on real datasets:

1. One variable
   - `CUBpq` is better or at least in line with other multivariate methods
   - little uncertainty and not strong relation with covariates → the median may be the best method
   - median produces more biased estimators
2. Likert structure, more variables
   - `CUBpq` stands on the performance of the other multivariate models
   - high uncertainty → CUBpq improves
   - MF appears the most performant and computationally efficient
   - MF produces more biased estimators